Human Language Technologies – The Baltic Perspective
A. Tavast et al. (Eds.)

doi:10.3233/978-1-61499-133-5-36

Data Pre-Processing to Train a Better Lithuanian-English MT System

Daiga DEKSNE[1] and Raivis SKADIŅŠ[2]
TILDE, Latvia

Abstract. In this paper, we present the results of a series of experiments done to improve the quality of a Lithuanian-English statistical MT (SMT) system. We particularly focus on word alignment and out of vocabulary issues in SMT translating from a morphologically rich language into English.

Keywords. statistical MT, data pre-processing, Lithuanian-English

Introduction

Lithuanian is a highly inflected language. The words change the form according to grammatical function. That means that the endings of nouns, pronouns, adjectives, numerals, and verbs change depending on certain features; and a word may have many different surface forms depending on its role in a sentence. English instead does not have such a rich feature system.

This difference between languages significantly impacts word and phrase alignment when training an SMT system. Typically one or two forms of an English noun have to be aligned to more than ten different surface forms of a corresponding Lithuanian noun. Similarly, English verb forms have to be aligned with many surface forms of the Lithuanian verb; and Lithuanian verbs have prefixes indicating negation and other semantic features while English verbs do not have prefixes and such information is expressed using modifying words.

Some word forms in the corpus used to train a SMT system are not as common as others, therefore a Lithuanian-English SMT system does not translate all word forms equally well. It is very common to get many out of vocabulary words when translating from Lithuanian into English.

[1] Corresponding author: TILDE, Vienības gatve 75a, Riga, Latvia; E-mail: daiga.deksne@tilde.lv
[2] Corresponding author: TILDE, Vienības gatve 75a, Riga, Latvia; E-mail: raivis.skadins@tilde.lv

1. Experiments

Four different experiments using the DGT-TM parallel corpus[3] [1] (~806,000 parallel sentences, ~703,000 monolingual sentences) where performed to train the MT system which performs better also on not so common word forms. Results where compared to the baseline SMT system trained on the original DGT corpus without any data pre-processing. Both baseline SMT system and SMT systems trained on pre-processed data were trained using the LetsMT! platform [2] which is based on the Moses SMT toolkit [3]. In all the experiments the corpus was transformed using finite state transducers.

1.1. The first experiment – prefixes and endings as separate tokens (System #1).

In this experiment, we apply several transformation rules to a Lithuanian text corpus. If possible, endings and prefixes are separated from word stems.

A list of non-inflected part of speech words is included in transducer. The words from the text corpus which are in this list are not changed. A list of prefixes and a list of endings are included in the transducer. Combination of an optional prefix from the prefixes list, a stem and an optional ending from endings list form the word. The stem can be any sequence of letters which is at least two symbols long. The transformed word is in form 'prefix- stem -ending'. The symbol '-' is used to signify that token is a prefix or an ending.

For example, a sentence (1) has a single non-inflected part of speech word "ir". Other words in this sentence belong to different inflected part of speech classes, the prefixes and/or the endings are separated from them in transformation process. The transformed text (2) has endings "-as", "-a", "-o", "-imo", "-is" and an prefix "ne-" as a separate tokens.

Priedas ir Protokolas yra neatskiriama šio Susitarimo dalis. (1)

Pried -as ir Protokol -as yr -a ne- atskiriam -a ši -o Susitar -imo dal -is. (2)

1.2. The second experiment – prefixes separated, endings replaced by tense and number feature values (System #2).

In this experiment, prefixes are separated from word stems, but endings are replaced by tense and number feature values which the particular ending represents. Transformed word is in form 'prefix- stem&featurevalues'. The same ending can symbolize several feature values. Some examples of feature value tags and their meaning – PRESPAST (present or past tense), SG (singular), PL (plural), PLPRES (plural number or present tense). If some particular ending can be both – singular and plural form ending – number feature value is not used. A list of non-inflected part of speech words and a list of personal pronouns are included in transducer. The words from the text corpus which are in these lists will not be changed.

In the transformed sentence (3), the ending "imo" is replaced by "SGPRES" tag as it can represent two feature values – singularity and present tense. The "PRES" tag has replaced endings "as", "a", "is". The several part of speech words can have the same

[3] Release 2007

endings, but a set of features for a different part of speech classes is different. A tense feature characterizes verbs, but not nouns. In the sample sentence (1), only the word "yra" is a verb.

Pried&PRES ir Protokol&PRES yr&PRES ne- atskiriam&PRES ši&PRESPAST Susitar&SGPRES dal&PRES. (3)

There is no distinction between verb stems and other stems in transducer. This leads to situation when tense feature values are also assigned to noun stems. Table 1 shows that only two singular forms have a tag SG and only two plural forms have a tag PL. In this example, we can see than for the first paradigm nouns the translation quality might improve only for the dative and locative case forms.

Table 1. System #2. Feature values assigned to different forms of the first paradigm nouns (the form's ending in brackets)

	Singular	Plural
Nominative	PRES (as)	PRESPAST (ai)
Genitive	PRESPAST (o)	- (ų)
Dative	SG (ui)	PL (ams)
Accusative	PRES (ą)	PAST (us)
Instrumental	PRES (u)	PRESPAST (ais)
Locative	SG (e)	PL (uose)

1.3. The third experiment – prefixes separated, all endings replaced by number feature values and verb endings also by time feature values (System #3).

In this experiment, the two lists of endings – the verb endings and the other endings – are used to avoid the drawbacks of System #2. The tense feature is applied only to verb endings. Transducer has a full list of verb stems for which the verb endings are allowed. Other endings are allowed to any two or more letter combination which is not in the verb stem list. The verb stems are with a higher weight than other stems. Same as before, a list of non-inflected part of speech words and a list of personal pronouns is used in the transducer.

Pried& ir Protokol& yr&PRES ne- atskiriam&SG ši&PRESPAST Susitar&SG dal&PRES. (4)

In the transformed sentence (4), the situation has improved, the two first paradigm nouns "priedas" and "protokolas" do not have the tense feature values. But the noun stem "dal" still has the tense feature values tag as this stem also can be a verb stem.

Table 2. System #3. Feature values assigned to different forms of the first paradigm nouns (the form's ending in brackets)

	Singular	Plural
Nominative	- (as)	- (ai)
Genitive	SG (o)	- (ų)
Dative	SG (ui)	PL (ams)
Accusative	SG (ą)	- (us)
Instrumental	- (u)	PL (ais)
Locative	SG (e)	PL (uose)

1.4. The fourth experiment – prefixes separated, endings deleted (System #4).

In this experiment, prefixes are separated as in the previous experiments, but endings are deleted. The transformed sentence (5) contains only the stems of the inflected part of speech words.

Pried ir Protokol yr ne- atskiriam ši Susitar dal. (5)

2. Evaluation

The baseline system and all the systems with pre-processed data where evaluated on a random subset of 1,000 sentences from the DGT-TM corpus using BLEU [4], NIST [5], TER [6] and METEOR [7] automatic metrics. Sentences from the evaluation corpus were not included in training data. Evaluation results are shown in Table 3.

Table 3. Evaluation scores on a random subset of the DGT-TM corpus

	BLEU	NIST	TER	METEOR
Baseline	49.04	9.2774	48.54	0.4214
System #1	47.53	9.1871	50.02	0.4199
System #2	49.17	9.2546	49.07	0.4208
System #3	**49.22**	**9.2886**	**48.28**	**0.4241**
System #4	47.99	9.1072	49.83	0.4186

Although the systems have been trained on the legislation domain DGT-TM corpus, we also evaluated them on a general balanced test corpus consisting of 512 sentences (the corpus contains diverse texts from legal domain, user manuals, news texts, fiction, etc.). See evaluation results in Table 4. System #4 with the prefixes separated and the endings deleted has the lowest BLEU score on both evaluation sets. System #2 with the prefixes separated and the time/number feature values performs slightly better. System #1 with separated prefixes and endings performs better than System #2. Only one system exceeds the baseline system — System #3 with prefixes separated and the number feature values and the tense feature values for verbs.

Table 4. Evaluation scores on balanced evaluation corpus (512 sentences)

	BLEU	NIST
Baseline	15.14	4.9721
System #1	14.92	4.9487
System #2	13.69	4.7170
System #3	15.38	4.9600
System #4	13.72	4.6859

3. Large-scale experiment

The results obtained using the DGT-TM corpus show that a better MT system can be built by applying different pre-processing methods to training data.

The next step was to train a Lithuanian-English SMT system on a larger, more general corpus. The parallel corpus contained 5.3 M sentences, the monolingual corpus contained 81 M sentences. We trained the baseline system without data pre-processing and the system with data pre-processing as in System #3. We chose the data pre-processing technique which showed the best results on a smaller corpus. We trained only one system on a bigger corpus with data pre-processing as SMT system training is a time consuming process.

Table 5. Evaluation scores for the large-scale Lithuanian-English system, evaluated on balanced evaluation set

	BLEU
Baseline	37.83
System with pre-processing	38.42

In Table 5, we can see that the system with pre-processed data outperforms the baseline system by 0.59 BLEU points. Human evaluation was also performed to compare both systems. For evaluation of the system, the same methodology was used as in [8]. Nine (9) human evaluators where asked to give preference to translation of the first or of the second system. The translations where presented at a random order. The results of the system with pre-processed data are slightly better, in 50.99% (±3.55%) of cases human evaluators judged its output to be better than the baseline system's output. However, evaluation results are not sufficient to say with strong confidence that the system with data pre-processing is better than the baseline system, because the difference between the systems is not statistically significant (50.99% - 3.55% < 50%).

4. Conclusions and the next steps

Experiments reported in this paper show that it is possible to improve the quality of SMT translation from a highly inflected language into English by pre-processing the training data. Even a simple method described in this paper gives significant improvement in SMT systems trained on a relatively small corpus and a large corpus.

This paper describes only 4 simple ways of data pre-processing. More sophisticated pre-processing should be tried in our next experiments. We are considering the use of more advanced tools such as part of speech tagger or morphological analyzer instead of finite state transducers with very limited lexicon. Experiments with different length of prefixes and suffixes are also considered.

Acknowledgements

The research within the project LetsMT! leading to these results has received funding from the ICT Policy Support Programme (ICT PSP), Theme 5 – Multilingual web, grant agreement n° 250456.

References

[1] Steinberger R., A. Eisele, S. Klocek, S. Pilos & P. Schlüter. DGT-TM: A freely Available Translation Memory in 22 Languages. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012)*, Istanbul, 21-27 May 2012.

[2] Vasiļjevs, A., Skadiņš, R., & Tiedemann, J. (2011b). LetsMT!: Cloud-Based Platform for Building User Tailored Machine Translation Engines. In *Proceedings of the 13th Machine Translation Summit* (pp. 507-511). Xiamen, China.

[3] Koehn, P., Federico, M., Cowan, B., Zens, R., Duer, C., Bojar, O., Constantin, A., Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation, *Proceedings of the ACL 2007 Demo and Poster Sessions* (pp. 177-180), Prague

[4] Papineni, K., Roukos, S., Ward, T., Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics. :* ACL

[5] Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of HLT-02*

[6] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas*

[7] Banerjee, S. & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005)*. Ann Arbor, Michigan

[8] Skadins, R., Goba, K., & Šics, V. Improving SMT for Baltic Languages with Factored Models*, Baltic HLT 2010*, October 7-8, 2010, Riga, Latvia